A Global Haplotype Analysis of the Myotonic Dystrophy Locus: Implications for the Evolution of Modern Humans and for the Origin of Myotonic Dystrophy Mutations

S. A. Tishkoff,[1,*] A. Goldman,[2] F. Calafell,[1] W. C. Speed,[1] A. S. Deinard,[1] B. Bonne-Tamir,[3] J. R. Kidd,[1] A. J. Pakstis,[1] T. Jenkins,[2] and K. K. Kidd[1]

[1]Department of Genetics, Yale University School of Medicine, New Haven; [2]Department of Human Genetics, School of Pathology, South African Institute for Medical Research, University of the Witwatersrand, Johannesburg; and [3]Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv, Tel Aviv

Summary

Haplotypes consisting of the $(CTG)_n$ repeat, as well as several flanking markers at the myotonic dystrophy (DM) locus, were analyzed in normal individuals from 25 human populations (5 African, 2 Middle Eastern, 3 European, 6 East Asian, 3 Pacific/Australo-Melanesian, and 6 Amerindian) and in five nonhuman primate species. Non-African populations have a subset of the haplotype diversity present in Africa, as well as a shared pattern of allelic association. $(CTG)_{18-35}$ alleles (large normal) were observed only in northeastern African and non-African populations and exhibit strong linkage disequilibrium with three markers flanking the $(CTG)_n$ repeat. The pattern of haplotype diversity and linkage disequilibrium observed supports a recent African-origin model of modern human evolution and suggests that the original mutation event that gave rise to DM-causing alleles arose in a population ancestral to non-Africans prior to migration of modern humans out of Africa.

Introduction

Myotonic dystrophy (DM [MIM 160900]) is an autosomal dominant multiorgan disorder characterized by muscle wasting and weakness, ptosis, myotonia, cataract, cardiomyopathy, gonadal atrophy, and, sometimes, mental deficiency (Harper 1989). DM is considered to be the most prevalent inherited neuromuscular disease of adults, with an incidence of 2.2–5.5/100,000 in western Europeans (Harper 1989) and 5.5/100,000 in Japanese (Davies et al. 1992). DM is less prevalent in Southeast Asians (Ashizawa and Epstein 1991) and is rare or absent in southern and central Africans, with only one case reported, in a Nigerian kindred (Dada 1973; Ashizawa and Epstein 1991; Krahe et al. 1995*b*). DM is one of several genetic disorders exhibiting genetic anticipation in families (an increase in severity and an earlier age at onset, in successive generations), because of expansion of unstable trinucleotide repeats (Aslanidis et al. 1992; Buxton et al. 1992; Harley et al. 1992, 1993; Hunter et al. 1992; Tsilfidis et al. 1992). Individuals affected with DM have an expansion of a $(CTG)_n$ repeat array in the 3′ UTR of the myotonin protein kinase gene (DMPK) on chromosome 19q13.3 (Brook et al. 1992; Fu et al. 1992). On normal chromosomes, the array of tandemly repeated CTGs is highly polymorphic, with a copy-number range of 5–38, whereas, on DM chromosomes, the copy number is >42 and has been found to be as high as 2,000 in severely affected individuals (Brook et al. 1992; Fu et al. 1992; Harley et al. 1992; Mahadevan et al. 1992). Such expanded arrays have been demonstrated to account for >99% of DM cases in individuals of diverse ethnic backgrounds (Harley et al. 1992; Mahadevan et al. 1992; Mulley et al. 1993; Yamagata et al. 1996). However, it is not yet clear whether DM is the result of altered expression or function of the DMPK protein or is the result of altered expression of genes neighboring the DMPK locus (Boucher et al. 1995; Jansen et al. 1995, 1996; Krahe et al. 1995*a;* Wang and Griffith 1995; Harris et al. 1996).

Haplotype analyses have demonstrated a complete allelic association between DM-causing alleles and a full-length allele, *Alu*(+), at a site with an *Alu*-deletion polymorphism located 5 kb telomeric to the $(CTG)_n$ repeat in all patients of European and Asian ancestry examined to date (Harley et al. 1992; Imbert et al. 1993; Mahadevan et al. 1993; Neville et al. 1994; Goldman et al. 1996*a*). Haplotype analyses incorporating as many as

Received September 10, 1997; accepted for publication March 19, 1998; electronically published May 20, 1998.

Address for correspondence and reprints: Dr. Kenneth K. Kidd, Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06520-8005. E-mail: kidd@biomed.med.yale.edu

[*] Present affiliation: Department of Biology, Pennsylvania State University, University Park.

0002-9297/98/6206-0016$02.00

eight additional polymorphisms spanning a physical distance of 30 kb have also shown complete association between DM-causing alleles with a single haplotype in individuals of European ancestry (Neville et al. 1994; Goldman et al. 1996*a*). There is strong, but not complete, linkage disequilibrium between DM-causing alleles and several extragenic polymorphisms extending $\leqslant$160 kb from the $(CTG)_n$ array (Imbert et al. 1993; Neville et al. 1994; Goldman et al. 1996*a*). On the basis of these observations, it was proposed that all DM-causing alleles have a common origin, possibly on a predisposing haplotype (Imbert et al. 1993; Mahadevan et al. 1993; Neville et al. 1994; Goldman et al. 1996*b*).

Initial studies of normal chromosomes from European and Japanese populations found a complete association between $(CTG)_5$ and $(CTG)_{\geqslant 18}$ alleles with the full-length *Alu* allele, *Alu*(+), and a nearly complete association between $(CTG)_{11–13}$ alleles and the 1-kb–deletion allele, *Alu*(-) (Yamagata et al. 1992; Imbert et al. 1993; Neville et al. 1994). These findings led to the hypothesis that large-sized normal alleles, $(CTG)_{\geqslant 18}$, initially arose from expansion of the $(CTG)_5$ allele on an *Alu*(+) haplotype background and now constitute a reservoir of unstable alleles that undergo recurrent expansion to the DM-causing state (Imbert et al. 1993; Neville et al. 1994). On the basis of the absence of large-sized alleles in sub-Saharan Africans, Goldman et. al. (1994, 1995, 1996*b*) proposed that the original DM mutation event occurred on an *Alu*(+) chromosome after the migration of modern humans out of Africa.

We have examined the distribution of $(CTG)_n$ alleles at the DMPK locus in normal individuals from 25 ethnically diverse human populations (5 African, 2 Middle Eastern, 3 European, 6 Asian, 3 Pacific/Australo-Melanesian, and 6 Amerindian) and in nonhuman primates and have extended the haplotype analysis to include several markers closely flanking the $(CTG)_n$ repeat. In addition to the much-studied *Alu*-deletion polymorphism, we chose to investigate systematically two other biallelic sites: a *Hin*fI site on which we already had some data and an easy-to-type *Taq*I site on the other side of the $(CTG)_n$ repeat. Our data indicate that the large-sized normal alleles arose in Africa from midsized alleles. A founder event, either before or concomitant with the migration out of Africa, established large-sized normal repeat alleles $(CTG)_{18–35}$ on a single haplotype background in the populations ancestral to modern non-African populations. Subsequent repeat-expansion events have produced DM-causing alleles from one or more of these large-sized normal alleles.

Subjects and Methods

Subjects

DNA samples from 25 populations were typed for this study of markers at the DM locus. All individuals sampled were normal, healthy volunteers not screened for DM status or for other disease. As it turned out, none of the $(CTG)_n$ repeat alleles observed in these samples is even large enough to be considered in the premutation range for DM. Most of the population samples have previously been described by Bowcock et al. (1991), Barr and Kidd (1993), Lichter et al. (1993), Castiglione et al. (1995), Goldman et al. (1996*b*), and Tishkoff et al. (1996*a,* 1996*b*). More-complete descriptions of many of these populations can be found on the Kidd lab homepage. Nearly all of these samples preexisted as Epstein-Barr virus–transformed lymphoblastoid cell lines (Anderson and Gusella 1984), and cell lines of 5–10 individuals of most populations are available through the Coriell Institute for Medical Research (National Institute of General Medical Sciences), Camden, N.J. All blood samples were obtained with informed consent, and these typings were done under protocols approved either by the Human Investigations Committee at Yale or by the Committee for Research on Human Subjects of the University of the Witwatersrand.

PCR Analyses

PCR analysis of $(CTG)_n$ repeats of the Bantu-speaking and !Kung San samples was performed as described elsewhere (Goldman et al. 1996*b*), by means of published primers 101 and 102 (Brook et al. 1992). All other samples were amplified by use of 50–100 ng of genomic DNA in a 25-μl (total volume) reaction mixture containing 20 pmol of fluorescent-labeled primer 101 and 20 pmol of unlabeled primer 102, 200 μM of each dNTP, 50 mM KCl, 10 mM Tris-HCl (pH 8.4), 1.5 mM $MgCl_2$, and .625 U of *Taq* polymerase (Perkin Elmer Cetus). Amplification products were run on an 8% polyacrylamide gel on an ABI 373A DNA sequencing machine, and fragment sizes were determined by Genescan software.

The *Alu* deletion was typed by use of the published primers (405, 486, and 491) and the published three-primer protocol (Mahadevan et al. 1993). Genomic DNA (100 ng) was PCR amplified with a mixture of 25 ng of primer 405, 25 ng of primer 491, 50 ng of primer 486, 1.5 mM $MgCl_2$, 500 mM KCl, 100 mM Tris-HCl, pH 8.3, 200 μM of each dNTP, and 1 U of *Taq* DNA polymerase, in a 25-μl reaction volume, with denaturation at 94°C for 5 min, followed by 30 cycles of 94°C for 1 min, 60°C for 1.5 min, and 72°C for 1.5 min.

The *Hin*fI polymorphism was typed by use of the published primers, Hh71 and Hh72, and the published protocol (Goldman et al. 1995). One hundred nanograms of genomic DNA was PCR amplified by use of 25 pmol of each primer, 1.5 mM $MgCl_2$, 500 mM KCl, 100 mM Tris-HCl, pH 8.3, 200 μM of each dNTP, and 1 U of *Taq* DNA polymerase, in a 25-μl reaction volume, with a 5-min denaturation at 94°C, followed by 30 cycles of 94°C for 1 min, 60°C for 1 min, and 72°C for 1 min.

The *Taq*I polymorphism was typed by use of the published primers, 565 and 564, and the protocol of Neville et al. (1994) but with the modifications described by Goldman et al. (1996*a*). One hundred nanograms of genomic DNA was PCR amplified by use of 25 pmol of each primer, 1.5 mM $MgCl_2$, 500 mM KCl, 100 mM Tris-HCl, pH 8.3, 200 μM of each dNTP, and 1 U of *Taq* DNA polymerase, in a 25-μl reaction volume, with a 5-min denaturation at 94°C, followed by 10 cycles of 94°C for 1 min, 60°C for 1 min, and 72°C for 1 min, followed by 20 cycles of 94°C for 1 min, 55°C for 1 min, and 72°C for 1 min.

Allele and Haplotype Frequencies

The allele frequencies at the separate sites (*Alu*, *Hin*fI, *Taq*I, and $[CTG]_n$) were estimated by gene counting. The computer program HAPLO (Hawley and Kidd 1995) was used to generate maximum-likelihood estimates of the haplotype frequencies. Grouping of $(CTG)_n$ alleles into classes was performed after inspection of the haplotype frequency estimates based on the separate alleles at the contributing $(CTG)_n$ site.

Linkage Disequilibrium

Gametic disequilibrium was quantified by use of the standardized linkage-disequilibrium parameter D', devised by Lewontin (1964). In order to test the null hypothesis that the pairwise linkage-disequilibrium coefficient is 0, a χ^2 test statistic was computed (Weir 1996, eq. [3.10]). These linkage-disequilibrium estimates and test statistics, as well as other disequilibrium estimates (which are not reported here) were calculated by the computer program LINKD written by A. J. Pakstis. The LINKD source code, along with examples of input and output files, are available, via ftp, from the Yale Center for Medical Informatics Website.

Results

$(CTG)_n$ Allele Distributions

In DNA samples from 1,235 people from 25 ethnically diverse human populations, the $(CTG)_n$ allele size range is 5–35 repeats. Frequencies of $(CTG)_n$ allele size classes, for each population studied, are shown in table 1. No population showed a significant departure from Hardy-Weinberg proportions. Most populations have a bimodal distribution of allele sizes, with peaks at 5 repeats and between 9 and 17 repeats (midsized alleles). The Ethiopian Jews and many of the non-African populations have a third group of large-sized alleles, range 18–35 repeats. Globally, there are distinct antimodes at 7–8 repeats and at 18 repeats. We observe considerable interpopulation variability of $(CTG)_n$ allele class frequencies (χ^2= 458; 48 df; $P < .0001$). The $(CTG)_{9-17}$ size class is the most frequent (range 46%–100%) class of alleles in all populations. The $(CTG)_5$ allele is the most common single allele in the African Biaka and Bantu-speaking populations and in the Middle Eastern, European, and Southeast Asian populations. However, we find a midsized allele (11–13 repeats) to be the most common in 12 of the 24 populations. The $(CTG)_5$ allele exists at low (1%–8%) frequency in the Siberian Yakut, Cheyenne, Jemez Pueblo, and Maya and is absent in the New Guinea Highlanders and in the Rondonia Surui, Ticuna, and Karitiana of Brazil. These observations are consistent with other studies, which find the $(CTG)_5$ allele to be rare or absent in New Guinea Highlanders, Tibetans, and Amerindians (Zerylnick et al. 1995; Deka et al. 1996). Except for the occurrence of two chromosomes with $(CTG)_{21}$ and $(CTG)_{22}$ in South African Bantu-speaking individuals, large-sized alleles ($\geqslant$18 repeats) are present only in the Ethiopian Jews and in non-African populations. The frequency of $(CTG)_{\geqslant 18}$ alleles is highest in the Ethiopian Jews (10%), Yemenite Jews (15%), Danes (9%), Japanese (9%), Yakut (10%), Nasioi Melanesians (18%), and Highland New Guineans (10%). The largest allele sizes (32 and 35 repeats) were identified in the Ethiopian and Yemenite Jews. However, alleles with 18–29 repeats were found in populations from all geographic regions outside Africa.

CTG repeat–number alleles were examined by PCR in 36 nonhuman primates; the resulting frequencies are shown in table 2. The $(CTG)_n$ repeat is polymorphic in both common and pygmy chimpanzees, with alleles in the range of 7–19 repeats. The 12-repeat allele is most frequent (12 of 38 chromosomes) in common chimpanzees, and the 9-repeat allele is most common (7 of 10 chromosomes) in pygmy chimpanzees. Four alleles were identified in the three orangutans examined, including one 30-repeat allele. The five gorillas examined are monomorphic for the 7-repeat allele, and the four gibbons examined are monomorphic for a 4-repeat allele.

Haplotype Distributions

Samples of all 1,235 humans were typed for the $(CTG)_n$ repeat, as well as for three biallelic markers encompassing a distance of 21 kb. The biallelic markers include a 1-kb deletion of three *Alu* elements (Mahadevan et al. 1993), a *Hin*fI restriction-site polymorphism (RSP) (Mahadevan et al. 1993), and a *Taq*I RSP (Neville et al. 1994) (fig. 1). Highly significant frequency differences for alleles at these three biallelic markers were found across populations (for the *Alu* deletion polymorphism, χ^2 = 734, 24 df, $P < .00001$; for the *Hin*fI RSP, χ^2 = 340, 24 df, $P < .0001$; and, for the *Taq*I RSP, χ^2 = 707, 24 df, $P < .0001$). The *Alu*(+), *Hin*fI(+) (site

Table 1

Frequencies of (CTG)$_n$ Alleles in 25 Human Populations

Population (2*N*)	Frequency of (CTG)$_n$ Size Class				No. of Different Alleles	Heterozygosity[b]
	5	6–8	9–17	18–35 (Size or Range[a])		
African:						
Biaka (120)	.325	.025	.648	0 (NA)	11	.81 ± .04
Mbuti (92)	.108	0	.892	0 (NA)	7	.79 ± .04
Bantu-speakers (358)	.249	.073	.673	.005 (21–22)	13	.84 ± .02
!Kung San (126)	.103	.071	.824	0 (NA)	10	.81 ± .03
Ethiopian Jews (132)	.159	.045	.697	.100 (20–32)	16	.84 ± .03
European/Middle Eastern:						
Yemenite Jews (108)	.389	0	.463	.147 (20–35)	14	.78 ± .04
Druze (100)	.420	0	.540	.040 (21–24)	11	.75 ± .04
Danes (98)	.285	0	.621	.091 (21–25)	13	.83 ± .04
Roman Jews (54)	.407	.019	.574	0 (NA)	7	.74 ± .06
MixedEuropeans (78)	.372	.013	.563	.052 (19–29)	12	.78 ± .05
Asian:						
Kochari (36)	.250	0	.750	0 (NA)	5	.78 ± .07
Chinese (84)	.369	0	.632	0 (NA)	7	.73 ± .05
Japanese (100)	.190	0	.720	.090 (20–28)	11	.76 ± .04
Yakut (102)	.069	0	.834	.098 (21–24)	8	.74 ± .04
Atayal (84)	.453	0	.547	0 (NA)	6	.70 ± .05
Cambodians (48)	.438	0	.564	0 (NA)	8	.73 ± .06
Pacific/Australo- Melanesian:						
Micronesians (56)	.357	0	.643	0 (NA)	6	.76 ± .06
Nasioi (46)	.109	0	.721	.177 (21–28)	9	.83 ± .06
New Guineans (40)	0	0	.900	.100 (19–21)	9	.77 ± .01
North American:						
Cheyenne (102)	.029	0	.932	.040 (22–26)	10	.68 ± .05
Jemez Pueblo (86)	.082	0	.873	.047 (23–27)	9	.76 ± .05
Maya (102)	.099	0	.862	.040 (22–28)	8	.66 ± .05
South American:						
Rondonian Surui (86)	0	0	.965	.036 (21–28)	6	.72 ± .05
Ticuna (122)	0	0	.992	.008 (18)	5	.45 ± .05
Karitiana (102)	0	0	1.00	0 (NA)	3	.25 ± .04

[a] NA = not applicable (when large alleles in size class 18–35 do not occur).

[b] Calculated as $1 - \Sigma p_i^2$, where p_i is the individual allele frequency.

present), and *Taq*I(+) (site present) alleles are generally most frequent in Africa, decrease in frequency across Eurasia, and are rare or absent in Amerindians. Genotype frequencies of the three polymorphisms are close to predicted Hardy-Weinberg expectations. Only the *Alu*(+), *Hin*fI(+), and *Taq*I(+) alleles were present in the 48 chimpanzee chromosomes surveyed. Only the *Alu*(+) and *Hin*fI(+) alleles were detected in the 10 gorilla chromosomes surveyed; PCR using existing primers failed to amplify the region encompassing the *Taq*I polymorphism in gorilla.

Haplotype frequencies were estimated from phenotype data from all four polymorphic sites, as well as from the three stable biallelic markers alone. Each of the eight possible haplotypes defined by the three biallelic polymorphisms was observed in at least one sub-Saharan African population and in at least one non-African population (fig. 2). However, the (+++) and (−−−) haplotypes account for 97% of all chromosomes in the northeastern African populations and non-African populations. These two haplotypes exist at approximately equal frequencies in the Ethiopian Jews, Yemenite Jews, Druze, and Europeans. The (−−−) haplotype increases in frequency, from west to east across Eurasia and from north to south in the New World, until it is the only haplotype observed in the South American Ticuna and Rondonian Surui. In sub-Saharan Africa, the (+++) and (−−−) haplotypes account for only 76% of all chromosomes. Several haplotypes that are rare outside Africa exist at moderate to high frequency in some African populations. For example, (+−+) occurs at a frequency of 41% among the !Kung San and at a frequency of 6% among Bantu-speakers but occurs at lower frequencies in other African samples, whereas outside of Africa it ranged from 1% in Yemenites and Chinese to 3% among mixed Europeans. This haplotype displays especially large differences among the African populations, which may make it useful for studies of the relationships among those populations. In general, the sub-Saharan African populations show highly divergent

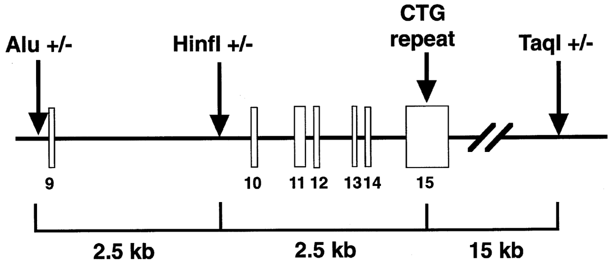

Figure 1 Schematic map of genomic DNA, showing exons 9–15 of the DMPK gene and the relative location of the *Alu*, *Hin*fI, and *Taq*I polymorphisms. "*Alu*(+)" denotes that the full-length *Alu* region is present; and "*Alu*(−)" denotes that it is absent; "*Hin*fI(+)" denotes that the *Hin*fI restriction site is present; and "*Taq*I(+)" denotes that the *Taq*I restriction site is present.

haplotype-frequency distributions ($\chi^2 = 210$, 21 df, $P <$ .0001; fig. 2).

Pairwise linkage disequilibrium (D') between the *Alu*, *Hin*fI, and *Taq*I polymorphisms is highly significant (P < .0001) in all populations (table 3). Compared with non-Africans, Sub-Saharan African populations have somewhat lower D' values—7 of 12 D' values are <.91 in the four sub-Saharan African populations, whereas none of the northeastern African populations or non-African populations have values <.93, and 14 of 20 non-African populations have D' values of 1.00, for all three pairwise comparisons. However, the D' values are significantly lower in sub-Saharan populations, as a group, only for the *Alu-Taq*I disequilibria: $P = .004$ for *Alu-Taq*I disequilibria, $P = .103$ for *Alu-Hin*fI disequilibria, and $P = .061$ for *Hin*fI-*Taq*I disequilibria (all values are based on the Mann-Whitney U-test).

Frequencies of four-site haplotypes, based on the three biallelic polymorphisms and the $(CTG)_n$ repeat, are shown graphically in figure 3, for 15 populations representative of the distributions seen in major world regions. Both the number of different haplotypes and the heterozygosity are highest in African populations, decrease from west to east across Eurasia and the Pacific, and are lowest in the Americas (table 4). In non-African populations, the (+++) haplotype is strongly associated with $(CTG)_5$ alleles (90% of all $[CTG]_5$ alleles), and the (−−−) haplotype is strongly associated with $(CTG)_{9–17}$ alleles (93% of all $[CTG]_{9–17}$ alleles) (table 5), in agreement with previous studies based on the *Alu* or the *Alu* and *Hin*fI polymorphisms in conjunction with the $(CTG)_n$ short tandem-repeat polymorphism (STRP) (Imbert et al. 1993; Neville et al. 1994). Similarly, in non-African populations and in Ethiopian Jews, the $(CTG)_{18–35}$ group of alleles is strongly associated with (+++) chromosomes (85%–92% of all $(CTG)_{18–35}$ alleles). The (+++) haplotype is also strongly associated with $(CTG)_5$ alleles in sub-Saharan African populations (95% of all $(CTG)_5$) and in Ethiopian Jews (94% of all $[CTG]_5$). However, the association between $(CTG)_{9–17}$ alleles and the seven other "background" haplotypes is strikingly different in the African populations: 46% of all $(CTG)_{9–17}$ alleles are associated with (+++) haplotypes in sub-Saharan Africans, compared with 21% in Ethiopian Jews; only 24% of all $(CTG)_{9–17}$ alleles are associated with (−−−) chromosomes in sub-Saharan Africans, compared with 74% in Ethiopian Jews; and 29% of all $(CTG)_{9–17}$ alleles are associated with other haplotypes in sub-Saharan Africans, compared with 6% in Ethiopian Jews.

Discussion

Global Frequency of $(CTG)_{\geqslant 18}$ Alleles

These data constitute the most extensive global survey to date of the frequency of $(CTG)_n$ alleles in normal individuals from 25 clearly defined, geographically and ethnically diverse populations. Our results agree with and extend previous studies demonstrating a correlation between the frequency of large-sized normal $(CTG)_n$ alleles and the prevalence of DM in different ethnic groups (Yamagata et al. 1992; Zerylnick et al. 1995; Deka et al. 1996; Goldman et al. 1995, 1996*a*). The frequencies of large-sized normal alleles outside Africa (table 1) correlate with the prevalence of DM, in that both are highest in western Europeans and Japanese (Harper 1989; Da-

Table 2

Observed Numbers of $(CTG)_n$ Alleles in Nonhuman Primates

	Nos. of $(CTG)_n$ Alleles for Repeat-Number Size =[a]													
Species	4	7	8	9	10	11	12	13	14	15	17	19	30	Total
Common chimpanzee	...	4	...	6	6	3	12	3	1	1	1	1	...	38
Pygmy chimpanzee	...	1	1	7	1	...	...	...	...	...	...	...	...	10
Gorilla	...	10	...	...	...	...	...	...	...	...	...	...	...	10
Orangutan	...	...	...	3	1	...	1	...	...	...	...	...	1	6
Gibbon	8	...	...	...	...	...	...	...	...	...	...	...	...	8

[a] Allele sizes are based on size comparability to humans. Chimpanzees, orangutans, and gibbons have variant sequences for one or two of the repeats (Rubinsztein et al. 1994).

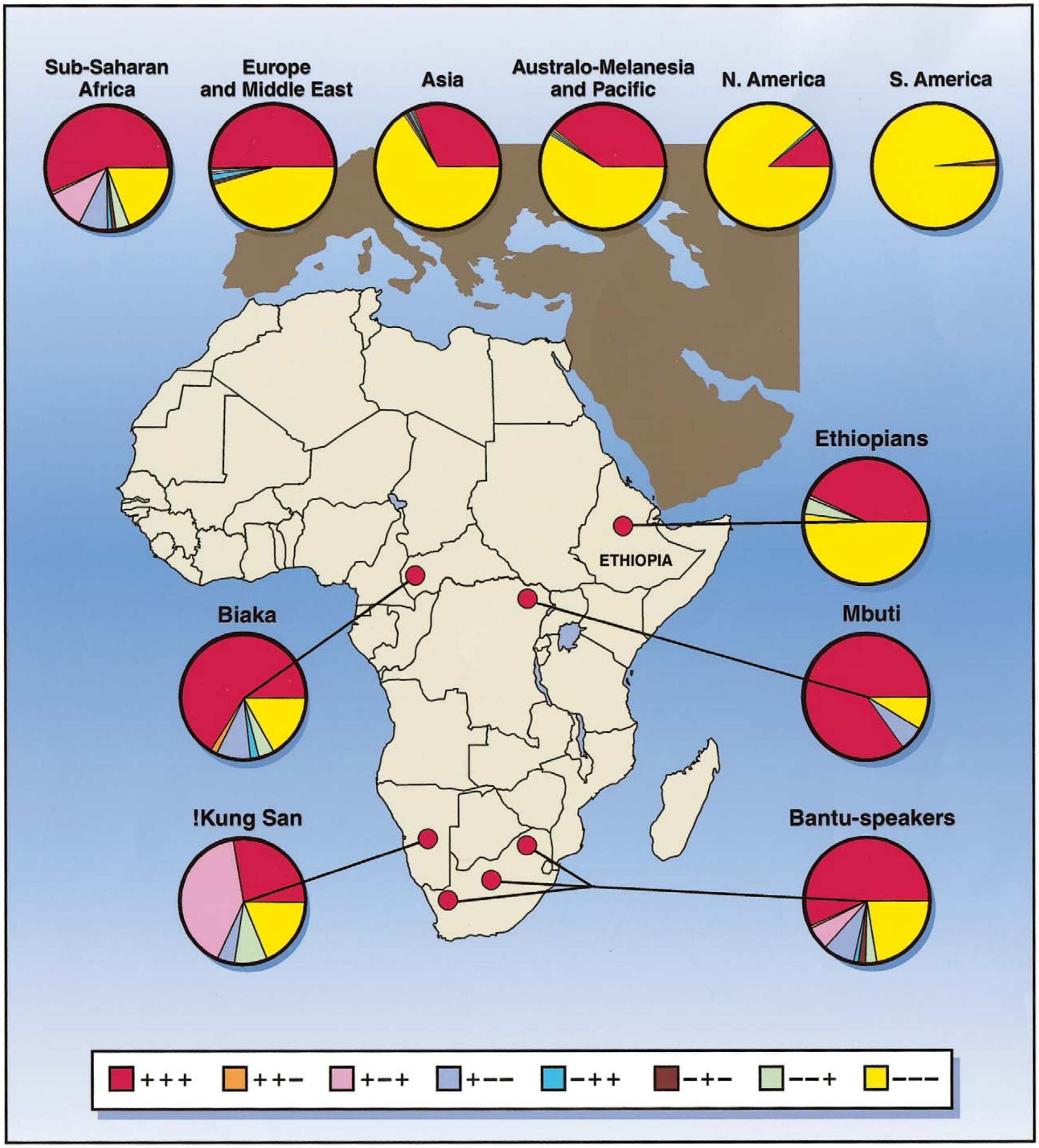

Figure 2 Background *Alu*/*Hin*fI/*Taq*I haplotype frequencies in African populations and in world populations, pooled by geographic region. Haplotypes are represented by the combination of allele symbols for the sites ordered as in figure 1; that is, (+++) haplotypes represent chromosomes containing the full-length *Alu* region and presence of the *Hin*fI and *Taq*I restriction sites. The color pie charts showing the haplotype frequencies are arranged counterclockwise, starting from three o'clock, in the order shown in the key, from left to right.

vies et al. 1992) and are lowest in Southeast Asians (Ashizawa and Epstein 1991). Yemenite Jews also have one of the highest reported prevalences of DM: 17/100,000, compared with an average of 2.2-5.5/100,000 in Europeans (Harper 1989; Korczyn 1994). We now find that they have a high (15%) frequency of $(CTG)_{\geqslant 18}$ alleles and the largest size range (5–35 repeats) of $(CTG)_n$ alleles seen in any of the populations that we have studied. This

Table 3

Pairwise Linkage Disequilibrium, *D′*, for *Alu, Hinf*I, and *Taq*I Polymorphisms

Population	*D′* Value for		
	*Alu-Hinf*I	*Hinf*I-*Taq*I	*Alu-Taq*I
African:			
Biaka	.86	.91	.61
Mbuti	1.00	1.00	1.00
!Kung San	1.00	.98	.77
Bantu-speakers	.80	.87	.77
Ethiopian Jews	1.00	1.00	1.00
Non-African:			
Yemenite Jews	.96	.96	.96
Mixed Europeans	.95	.94	1.00
Chinese	.94	.95	1.00
Atayal	1.00	.94	.94
Yakut	1.00	.93	1.00
Nasioi	.94	.93	.93
All 14 other populations	1.00	1.00	1.00

finding is consistent with a previous study of Yemenite Jews, in which alleles with 5–28 repeats were observed (Mor-Cohen et al. 1997*b*). $(CTG)_{\geqslant 18}$ alleles are rare among the four sub-Saharan African populations, which is consistent with the virtual absence of DM among southern and central African populations, with only one confirmed case, in a Nigerian family (Lotz and van den Meyden 1985; Ashizawa and Epstein 1991; Goldman et al. 1996*b*). However, the northeastern-African Ethiopian Jews have a high (10%) frequency of $(CTG)_{\geqslant 18}$ alleles, up to 32 repeats in size. This is compatible with previous reports (Mor-Cohen et al. 1997*a*, 1997*b*) of 5–26 repeats in Ethiopian Jews and of a high prevalence of DM in the population. These findings support the hypothesis that the high frequency of large-sized repeats in the normal range may result in a greater predisposition toward DM (Yamagata et al. 1992; Zerylnick et al. 1995; Goldman et al. 1995; Deka et al. 1996; Mor-Cohen et al. 1997*a*, 1997*b*).

Patterns of Haplotype Variation and Linkage Disequilibrium

From the presence of only the *Alu* (+), *Hinf*I (+), and *Taq*I (+) alleles in nonhuman primates, we infer that the (+++) haplotype is ancestral in humans. The 1-kb deletion of three *Alu* elements and the mutations at the *Hinf*I and *Taq*I sites most likely occurred after the divergence of humans from the great apes, 4–6 million years ago (Wilson and Sarich 1969; Hill 1994). The probability of any of these events occurring more than once during human evolution is quite low. The most common haplotypes in most populations are (+++) and (−−−). The sequence of events (three mutations and, possibly, recombination) leading to the triply derived haplotype (−−−) is not known. Similarly, the sequence of events producing any given copy of the other six haplotypes—that is, those with a combination of ancestral and derived alleles—is not known. In sub-Saharan African populations, the ancestral (+++) haplotype is most frequent. However, we observe several of the other haplotype combinations, such as (−−+), (+−−), and (+−+), at high frequency, suggesting that sub-Saharan African populations still retain these "intermediate" haplotypes and/or have had more time for new haplotypes to be formed by recombination among existing haplotypes. Genetic drift is likely responsible for the (+++) and (−−−) haplotypes having reached nearly equal high frequencies in the Ethiopian Jews and in the Middle Eastern and European populations, with a concomitant loss of nearly all other haplotypes. The (−−−) haplotype has drifted to higher frequency in Asia, the Pacific Islands, and the New World, until it is the only haplotype seen in the Brazilian Ticuna sample (fig. 2). This clinal pattern of variation is consistent with a series of founding events as modern humans emerged from Africa and migrated west to east across Eurasia during the past 100,000 years (Cavalli-Sforza et al. 1994, pp. 154–157) and north to south across the New World during the past 15,000–30,000 years (Dillehay 1989, p. 306; Adovasio et al. 1978).

The overall pattern of diversity of haplotypes incorporating the $(CTG)_n$ repeat (fig. 3) reflects both the diversity of $(CTG)_n$ alleles on each background haplotype and the background haplotype diversity. We observe both a high level of haplotype diversity in African populations and a subset of that diversity, as well as a shared pattern of allelic association, in non-African populations. Invoking selection to explain this pattern would require a very specific type of ad hoc selection for which we have no other evidence. This pattern of haplotype variation and linkage disequilibrium is consistent with the pattern of variation observed at the CD4 locus (Tishkoff et al. 1996*b*) and the DRD2 locus (Castiglione et al. 1995; Kidd et al. 1996, and in press). This global pattern can best be explained by a recent common origin of the non-African populations, from a population with reduced genetic variation relative to its African progenitor (possibly a northeastern-African population), and subsequent genetic drift as modern humans spread around the rest of the globe and diversified. Thus, these data are consistent with the "out of Africa" hypothesis of modern human origins, which purports that all human populations descend from an anatomically modern *Homo sapiens* population that migrated out of Africa ~100,000 years ago (Stringer and Andrews 1988).

In non-African populations, both the association of (+++) chromosomes with $(CTG)_5$ and $(CTG)_{\geqslant 18}$ alleles and the association of (−−−) chromosomes with $(CTG)_{9–17}$ alleles are strong but not complete. The (+++) chromosomes containing $(CTG)_{9–17}$ repeats,

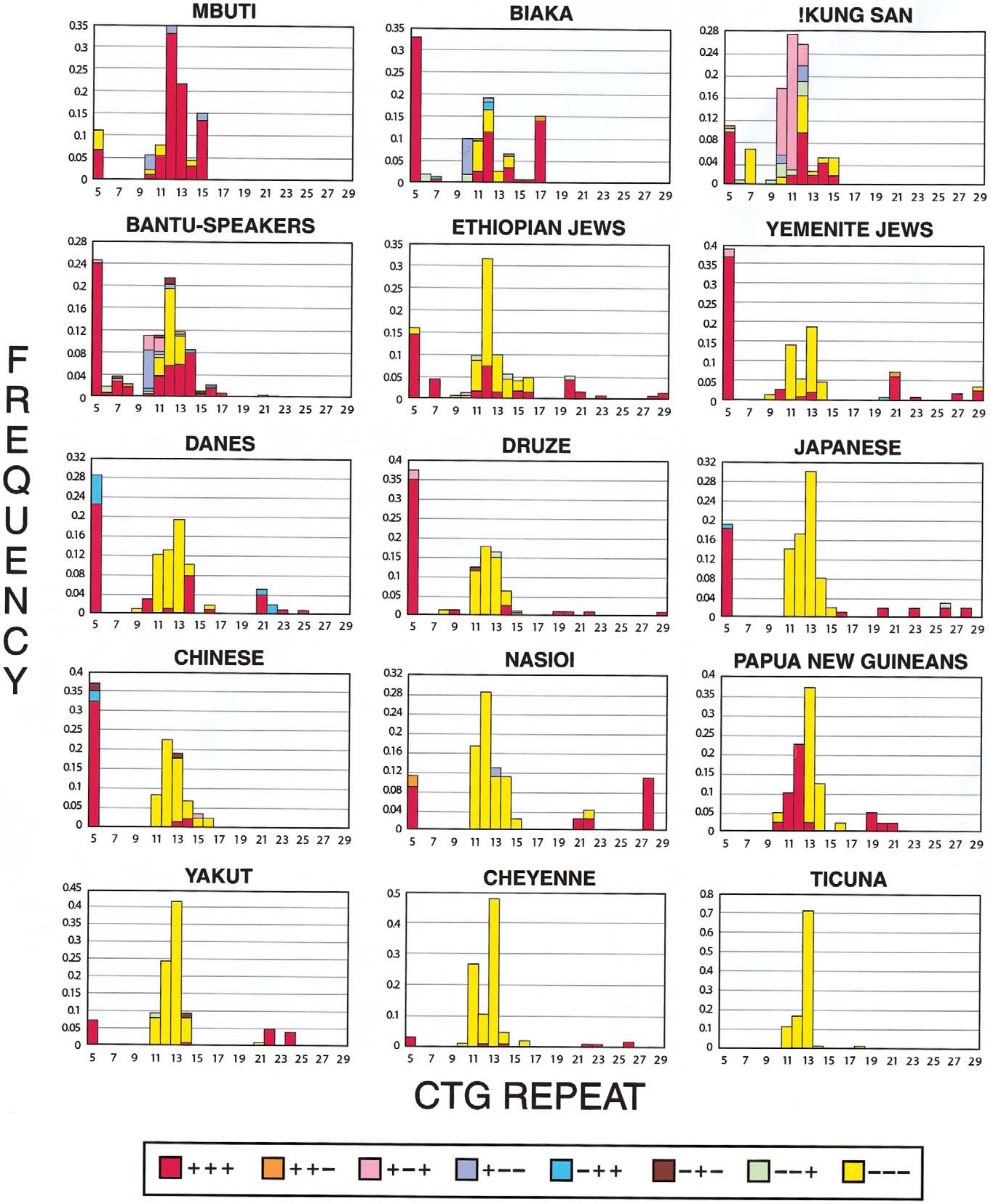

Figure 3 Frequencies of haplotypes of all four polymorphisms in a globally distributed set of population samples. Haplotypes are graphed according to the number of $(CTG)_n$ repeats, by means of stacked bars to represent the background haplotypes. $(CTG)_n$ alleles >29 repeats in the Ethiopian and Yemenite Jews have been pooled with $(CTG)_{29}$ alleles. Note that different frequency scales are used for the different populations. Population samples not shown had distributions very similar to those geographically closest populations that are shown.

Table 4

Haplotype Diversity, by Geographic Region

Region (No. of Populations)	Sample Size (2*N*)	No. of Haplotypes		Mean of Individual Population Heterozygosities[c]
		Total[a]	Mean (Range)/Population[b]	
Africa (5)	828	58	24.8 (14–41)	.87 ± .02
Europe/Middle East (5)	438	39	15.6 (9–19)	.80 ± .02
Asia (6)	454	27	10.1 (6–13)	.77 ± .02
Pacific Islands/Australo-Melanesia (3)	150	21	10.7 (9–12)	.81 ± .02
North America (3)	290	22	11.7 (11–12)	.71 ± .03
South America (3)	310	10	5.3 (4– 7)	.48 ± .13

[a] Data are based on the pooled data for each geographic region; haplotypes included the $(CTG)_n$ repeat as well as the flanking *Alu*(+/−), *Hin*fI, and *Taq*I polymorphisms.

[b] Data are estimated as the unweighted mean for the population.

[c] Data represent the unweighted means of expected haplotype heterozygosities for each population in a region.

which exist at high frequency in Ethiopian Jews, may have existed at low frequency in the ancestral population that migrated from Africa and may have drifted to a higher frequency in some of the non-African populations (e.g., Danes and New Guineans). Outside Africa, almost 5% of $(CTG)_5$ alleles and 8.7% of $(CTG)_{\geqslant 18}$ alleles occur on haplotype backgrounds that could have originated by rare single-recombination events between (+++) chromosomes and (−−−) chromosomes (table 5). However, in the case of $(CTG)_5$ and $(CTG)_{\geqslant 18}$ alleles on (−−−) chromosomes, the close proximity of the biallelic markers makes it unlikely that multiple recombination events could account for the transition from (+++). These rare haplotypes could have resulted either from "gene conversion" events, such as unequal sister-chromatid recombination, on (−−−) chromosomes or by means of rare mutational transits from the $(CTG)_{9–17}$ group of alleles on a (−−−) background. Therefore, it seems likely that the only chromosomes with $(CTG)_5$ and $(CTG)_{\geqslant 18}$ alleles in the ancestral non-African populations were on (+++) haplotype backgrounds and that all other haplotype combinations in non-Africans arose by recombination or mutation events after the migration of modern humans out of Africa.

Evolution of the $(CTG)_n$ Polymorphism

On the basis of initial haplotype findings in Europeans, it had been proposed that midsized alleles, $(CTG)_{9–13}$, arose on an *Alu*(−) chromosome and that $(CTG)_{\geqslant 18}$ alleles arose by a rare expansion event from a $(CTG)_5$ allele on an *Alu*(+) chromosome (Imbert et al. 1993). More-recent studies of a larger number of ethnically diverse populations have found that midsized alleles are sometimes associated with *Alu*(+) chromosomes (particularly in African populations) and that $(CTG)_5$ and $(CTG)_{\geqslant 18}$ alleles are occasionally observed on *Alu*(−) chromosomes (Rubinsztein et al. 1994; Goldman et al. 1995; Zerylnick et al. 1995; Deka et al. 1996). These studies suggested that the $(CTG)_{\geqslant 18}$ alleles may have arisen from the $(CTG)_{9–13}$ alleles by small changes in repeat size, rather than by a large "jump" from a $(CTG)_5$ allele. On the basis of the findings that nonhuman primates have only the full-length *Alu* and that the $(CTG)_5$-*Alu*(+) haplotype is most common in human populations, several authors have proposed that the $(CTG)_5$-*Alu*(+) haplotype is ancestral in humans (Imbert et al. 1993; Rubinsztein et al. 1994; Deka et al. 1996).

Our more extensive data lead to somewhat different conclusions. $(CTG)_n$ alleles having fewer than 7 repeats have not been seen in any great ape, although smaller alleles are seen in the lesser apes (gibbons). We find that midsized alleles (7–13 repeats) are most common in chimpanzees (table 2) and that a repeat allele in the 11–13 range is the most frequent allele in 12 of the 24 human populations examined. Furthermore, in sub-Saharan Africans, $(CTG)_{9–17}$ alleles are associated with many different haplotype backgrounds. Thus, it seems most likely that the ancestral (+++) haplotype in humans contained a mid-sized repeat and that the $(CTG)_5$ allele arose more recently, from a $(CTG)_{9–17}$ allele on a (+++) chromosome, by small changes in repeat size (we observed alleles with 6–8 repeats on (+++) chromosomes in several sub-Saharan African populations). However, the high frequency of $(CTG)_7$ repeats in chimpanzees and gorillas implies that the $(CTG)_7$ repeat is not inherently unstable. (We have not sequenced that allele in those apes; it might contain a stabilizing mutation.) The high frequency of $(CTG)_5$-(+++) haplotypes in many populations could be due to recent effects of random genetic drift, to increased mutational stability of the $(CTG)_5$ allele, and/or to selective or mechanistic constraint against mutation to alleles with <5 repeats. The other seven background haplotypes must have arisen more recently, through one or more mutation and/

Table 5

Distributions of Background Haplotypes for Each $(CTG)_n$ Allele-Size Class, by Geographic Region

	Frequency of $(CTG)_n$ Allele-Size Class (%)			
Population Group and Haplotype[a]	5 Repeats	9–17 Repeats	$\geq$18 Repeats	Total
Sub-Saharan Africans (4 populations):				
(+++)	94.91	46.07	0	56.75
(++−)	0	.64	0	.65
(+−+)	1.64	13.85	0	10.42
(+−−)	0	10.34	0	7.51
(−++)	0	1.29	0	.93
(−+−)	0	1.09	0	.93
(−−+)	1.25	2.82	100	3.74
(−−−)	4.00	23.90	0	19.07
Total	100	100	100	100
	(*N* = 151)	(*N* = 505)	(*N* = 2)	(*N* = 658)
Northeastern Africans (Ethiopian Jews):				
(+++)	94.05	20.71	92.45	43.16
(++−)	0	0	0	0
(+−+)	0	1.09	0	.80
(+−−)	0	0	0	0
(−++)	0	0	0	0
(−+−)	0	0	0	0
(−−+)	0	4.63	7.55	3.99
(−−−)	5.95	73.57	0	52.05
Total	100	100	100	100
	(*N* = 21)	(*N* = 92)	(*N* = 13)	(*N* = 126)
Non-Africans (20 populations):				
(+++)	91.94	5.90	84.96	27.24
(++−)	1.24	0	2.51	.38
(+−+)	1.23	.08	0	.31
(+−−)	0	.29	0	.21
(−++)	3.61	.08	4.93	1.04
(−+−)	0.60	.42	0	.44
(−−+)	0	.26	1.26	.26
(−−−)	1.83	92.97	6.34	70.12
Total	100	100	100	100
	(*N* = 333)	(*N* = 1,187)	(*N* = 80)	(*N* = 1,600)

[a] The three sites represented within each haplotype include the *Alu*(+/−), *Hin*fI, and *Taq*I polymorphisms. Each group of populations shows significant overall heterogeneity, at $P < .0001$, by a likelihood-ratio χ^2 test. Frequency data for the separate populations within each region were pooled.

or recombination events, probably on chromosomes initially containing midsized repeats. Mutation, recombination, and drift have produced the diversity of haplotypes in sub-Saharan populations, with drift, including a significant founder effect, possibly being responsible for both the high frequency of the triply derived (−−−) haplotype and its strong association with midsized $(CTG)_n$ alleles, especially in the non-African populations. In most populations, the midsized $(CTG)_n$ alleles have a unimodal distribution, whether they are on (+++) or (−−−) haplotype backgrounds; this is consistent with a mutation model, in which expansion/contraction occurs in small steps of one or a few repeat sizes (Shriver et al. 1993; Valdes et al. 1993; DiRienzo et al. 1994). However, the presence of large-sized $(CTG)_n$ repeats on the (−−−) background haplotypes outside Africa suggests that large jumps in repeat size do occur occasionally.

Origin of $(CTG)_{\geq 18}$ Alleles

In the sub-Saharan African populations, (+++) chromosomes are frequently associated with $(CTG)_{9\text{–}17}$ alleles. In the Ethiopian Jewish population, we observe a nearly continuous distribution, range 5–32 repeats, of $(CTG)_n$ alleles on (+++) chromosomes. Thus, our data support the hypothesis that the $(CTG)_{\geq 18}$ class of alleles arose from the $(CTG)_{9\text{–}17}$ class of alleles, rather than by a large jump from a $(CTG)_5$ allele (Rubinsztein et al. 1994; Zerylnick et al. 1995; Deka et al. 1996). Except for a 21-repeat allele and a 22-repeat allele that exist, in Bantu-speakers, on chromosomes containing the *Alu*(−) allele, $(CTG)_n$ alleles >17 repeats are observed only in Ethiopian Jews and non-Africans. Expansion into the $(CTG)_{18-35}$ range on a (+++) chromosome may have occurred by accumulating small changes in repeat size from the $(CTG)_{9-17}$ class of alleles, by a large mutational jump, or by an unequal recombination event (including sister-chromatid exchange) between two (+++) chromosomes containing $(CTG)_{9-17}$ alleles. By whatever mechanism, the event seems to be unlikely a priori, on the basis of the rarity of "new" large-size alleles in modern sub-Saharan populations, although the observation of a $(CTG)_n$ allele containing 30 repeats in the orangutan indicates that mutations into the large-sized range of alleles are not unique to humans. The expansion of $(CTG)_n$ repeats into the large-sized range in humans appears to have originated in an ancestral northeastern-African population prior to the migration of modern humans out of Africa and to have then been brought out of Africa with that migration. This expansion of the $(CTG)_n$ alleles may have crossed a threshold level of repeat number beyond which mutation occurs at a higher rate, explaining the broad distribution of $(CTG)_{\geqslant 18}$ alleles outside Africa. Our data support the model proposed by Imbert et al. (1993)—that is, that the stability of $(CTG)_n$ alleles is dependent on their length: $(CTG)_5$ alleles are most stable, $(CTG)_{9-17}$ alleles less stable, and $(CTG)_{\geqslant 18}$ alleles are least stable. This hypothesis is consistent with recent reports that the stability of microsatellites decreases with increasing length (Weber and Wong 1993; Zhang et al. 1994; Primmer et al. 1996). These data also demonstrate that both the founder effect during migration of modern humans out of Africa and subsequent genetic drift can dramatically alter the distribution of STRP alleles and may obscure their mutational history. Thus, these data clearly demonstrate the need to be cautious about the use of the distribution of STRP alleles in non-African populations to infer the mutation mechanisms or mutation rates of these STRPs.

$(CTG)_n$ alleles with 50–80 copies, originally termed the DM "protomutation" by Barceló et al. (1993), have been observed in unaffected members of DM families, and alleles in this range can expand into full-size DM-causing alleles (Mahadevan et al. 1993). The haplotype distributions observed (table 5 and fig. 3) provide a global perspective on the origin of $(CTG)_n$ alleles in the DM protomutation range, since $(CTG)_{18-35}$ alleles are the likely source of $(CTG)_{50-80}$ protomutations (Imbert et al. 1993; Neville et al. 1994). On the basis of earlier studies of the distribution of normal $(CTG)_n$ alleles in a smaller sample of African and non-African populations, Goldman et al. (1995) proposed that the initial mutation event that gave rise to the pool of DM protomutation alleles occurred on an *Alu*(+) chromosome after the migration of modern humans out of Africa. Our data suggest that this transition may have occurred on a (+++) chromosome in an ancestral northeastern-African population prior to migration out of Africa.

Haplotype analyses of DM patients of European descent have demonstrated complete association between DM alleles and a single haplotype containing the *Alu*(+), *Hin*fI(+), and *Taq*I (+) alleles (Neville et al. 1994; Goldman et al. 1996*a*). Our finding of strong linkage disequilibrium between $(CTG)_{\geqslant 18}$ alleles and (+++) chromosomes in Ethiopian Jews and non-African populations lends support to the hypothesis that $(CTG)_{\geqslant 18}$ alleles form a pool of unstable alleles that may expand into the DM protomutation range (Imbert et al. 1993). If this is the case, however, this raises the question of why 8% and 15% of $(CTG)_{\geq 18}$ alleles occur on a haplotype background other than (+++), in northeastern Africans and non-Africans, respectively, whereas 100% of DM-causing $(CTG)_{\geqslant 50}$ alleles in non-Africans exist on a (+++) haplotype. There are several possible explanations for this observed pattern of association and for the origin of DM-causing alleles.

1. All DM mutations descend from one or a few founder chromosomes that originated from one or a few rare expansion events from a $(CTG)_{18-35}$ allele on a (+++) chromosome to a $(CTG)_{50-80}$ allele in the ancestral non-African population prior to population divergence.
2. DM alleles have arisen independently in different ethnic groups subsequent to divergence, by recurrent low-frequency mutation of $(CTG)_{18-35}$ alleles into a pool of DM protomutations (50–80 repeats), which have an even higher probability of mutating into the DM size range.
3. DM alleles exist on a specific haplotype background that predisposes $(CTG)_n$ alleles toward expansion.

To help resolve these possibilities, which are not mutually exclusive, it would be necessary to perform a much more extensive survey of the prevalence of DM in many different ethnic groups, in order to demonstrate how strong the positive correlation is between the frequency of $(CTG)_{18-35}$ alleles and DM; a strong correlation would support hypothesis 2. Our results suggest that Ethiopian Jewish, Amerindian, Micronesian, and Australo-Melanesian populations, which have a high frequency of $(CTG)_{\geqslant 18}$ alleles, would be of particular interest to study for the prevalence of DM. Hypothesis 2 also predicts that, as "background" variation accumulates on chromosomes with $(CTG)_{\geqslant 18}$ alleles in different populations, a pool of non-(+++) chromosomes containing DM protomutations will accumulate. The complete association observed between DM-causing alleles and (+++) chromosomes could be due either to the fact that there has not been enough time for DM-causing alleles on non-(+++) chromosomes to accumulate or to the fact that most DM haplotype studies have been done in families from populations that have undergone recent founding events; thus, these chromosomes would be more likely to be identical by descent. As more DM patients originating from diverse ethnic populations are examined, DM alleles on non-(+++) chromosomes may eventually be identified. One such de novo mutation of a DM allele, on a chromosome containing an *Alu*(−) allele, has already been identified in the single sub-Saharan African kindred in which DM has been detected (Krahe et al. 1995*a*).

Hypothesis 1 would appear to be contradicted by the results of Brunner et al. (1993) and studies reviewed therein, which show a rapid transition from <100 (mean 67) repeats to full-length mutations. However, the available data all derive from retrospective studies of alleles ascertained because they did expand, in at least one descendant, to cause DM; it is possible that initiation of that rapid expansion is still quite rare. Brunner et al. (1993, p. 1020) note that their data do not allow them "to estimate, in absolute figures, the risk that a small mutation [<100 repeats—which would result in very mild, if any, symptoms] will evolve into a large mutation [>100 repeats—which would result in clinically significant symptoms]."

Hypothesis 3 cannot be excluded, but the presence of $(CTG)_5$ alleles on (+++) chromosomes argues that it cannot be this particular haplotype that predisposes the $(CTG)_n$ in cis to expansion—and that possibly some unidentified variants in the flanking sequence of a subset of (+++) chromosomes in the Ethiopian Jews and non-Africans do so. Thus, it would be useful to look for sequence variants flanking the $(CTG)_n$ repeat and to extend the haplotype analysis to include additional markers. Initial studies of markers located 90–160 kb from the DM gene in European populations have demonstrated a shared pattern of disequilibrium between DM-causing alleles and $(CTG)_{\geqslant 18}$ alleles, adding support to the model in which the $(CTG)_{\geqslant 18}$ alleles represent a pool from which DM-causing mutations are or have been formed (Imbert et al. 1993; Neville et al. 1994; Goldman et al. 1996*a*). A broader population survey using the markers that we have studied and any other variants identified in the region may provide additional insight into both the evolution of this chromosomal region and the mutation process leading to disease-causing alleles.

Acknowledgments

This work was supported in part by U.S. Public Health Service grants MH39239, and MH50390 (to K.K.K.), by National Science Foundation grant SBR9632509 (to J.R.K.), by a South Africa Medical Research Council grant (to T.J.), by a CIRIT (Catalonia, Spain) postdoctoral fellowship (awarded to F.C.), and by a Sloan postdoctoral fellowship in molecular evolution (awarded to S.A.T.). Collection of human population samples used for these studies was supported in part by U.S. Public Health Service grant AA09379 and by an Alfred P. Sloan foundation grant (both to K.K.K. and J.R.K.). We want to thank the individuals who donated the samples that made this study possible, and we also want to thank the numerous colleagues who assisted us in assembling the collection of samples.

Electronic-Database Information

URLs for data in this article are as follows:

Kidd lab, http://info.med.yale.edu/genetics/kkidd (for populations)

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/omim (for myotonic dystrophy [MIM 160900])

Yale Center for Medical Informatics, ftp://paella.med.yale.edu (directory pub/haplo)

References

Adovasio JM, Gunn JD, Donahue J, Stukenrath R (1978) Meadowcraft rockshelter, 1977: an overview. Am Antiquity 43: 632–651

Anderson MA, Gusella J (1984) Use of cyclosporin A in establishing Epstein-Barr virus-transformed human lymphoblastoid cell lines. In Vitro 20:856–858

Ashizawa T, Epstein HF (1991) Ethnic distribution of myotonic dystrophy gene. Lancet 338:642–643

Aslanidis C, Jansen G, Amemiya C, Shutler G, Mahadevan M, Tsilfidis C, Chen C, et al (1992) Cloning of the essential myotonic dystrophy region and mapping of the putative defect. Nature 355:548–551

Barceló JM, Mahadevan MS, Tsilfidis C, MacKenzie AE, Korneluk RG (1993) Intergenerational stability of the myotonic dystrophy protomutation. Hum Mol Genet 2:705–709

Barr CL, Kidd KK (1993) Population frequencies of the A1 allele at the dopamine D_2 receptor locus. Biol Psychiatry 34: 204–209

Boucher CA, King SK, Carey N, Krahe R, Winchester CL, Rahman S, Creavin T, et al (1995) A novel homeodomain-encoding gene is associated with a large CpG island interrupted by the myotonic dystrophy unstable $(CTG)_n$ repeat. Hum Mol Genet 4:1919–1925

Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Cartenuto L, Kidd KK, Cavalli-Sforza LL (1991) Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Proc Natl Acad Sci USA 88: 839–843

Brook, JD, McCurrach ME, Harley HG, Buckler AJ, Church D, Aburatani H, Hunter K, et al (1992) Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein family member. Cell 68:799–808

Brunner HG, Bruggenwirth HT, Nillesen W, Jansen G, Hamel BCJ, Hoppe RLE, de Die CEM, et al (1993) Influence of sex of the transmitting parent as well as of parental allele size on the CTG expansion in myotonic dystrophy (DM). Am J Hum Genet 53:1016–1023

Buxton J, Shelbourne P, Davies J, Jones C, Van Tongeren T, Aslanidis C, de Jong P, et al (1992) Detection of an unstable fragment of DNA specific to individuals with myotonic dystrophy. Nature 355:547–548

Castiglione CM, Deinard AS, Speed WC, Sirugo G, Rosenbaum HC, Zhang Y, Grandy DK, et al (1995) Evolution of haplotypes at the DRD2 locus. Am J Hum Genet 57: 1445–1456

Cavalli-Sforza L, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton

Dada TO (1973) Dystrophia myotonica in Nigerian family. East Afr Med J 50:213–228

Davies JD, Yamagata H, Shelbourne P, Buxton J, Ogihara T, Nokelainen P, Nakagawa M, et al (1992) Comparison of the myotonic dystrophy associated CTG repeat in European and Japanese populations. J Med Genet 29:766–769

Deka R, Majumder PP, Shriver MD, Stivers DN, Zhong Y, Yu LM, Barrantes R, et al (1996) Distribution and evolution of CTG repeats in the myotonin protein kinase gene in human populations. Genome Res 6:142–154

Dillehay TD (1989) Monte Verde: a late Pleistocene settlement in Chile. Vol. 1: Paleoenvironment and site context. Smithsonian Institution, Washington, DC

DiRienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB (1994) Mutational processes of simple-sequence repeat loci in human populations. Proc Natl Acad Sci USA 91:3166–3170

Fu YH, Pizzuti A, Fenwick RG Jr, King J, Rajnarayan S, Dunne PW, Dubel J, et al (1992) An unstable triplet repeat in a gene related to myotonic muscular dystrophy. Science 255: 1256–1258

Goldman A, Krause A, Ramsay M, Jenkins T (1996*a*) Founder effect and the prevalence of myotonic dystrophy in South Africans: molecular studies. Am J Hum Genet 59:445–452

Goldman A, Ramsay M, Jenkins T (1994) Absence of myotonic dystrophy in southern African Negroids is associated with a significantly lower number of CTG trinucleotide repeats. J Med Genet 31:37–40

——— (1995) New founder haplotypes at the myotonic dystrophy locus in Southern Africa. Am J Hum Genet 56: 1373–1378

——— (1996*b*) Ethnicity and myotonic dystrophy: a possible explanation for its absence in sub-Saharan Africa. Ann Hum Genet 60:57–65

Harley HG, Brook JD, Rundle SA, Crow S, Reardon W, Buckler AJ, Harper PS, et al (1992) Expansion of an unstable DNA region and phenotypic variation in myotonic dystrophy. Nature 355:545–546

Harley HG, Rundle SA, MacMillan JC, Myring J, Brook JD, Crow S, Reardon W, et al (1993) Size of the unstable CTG repeat sequence in relation to phenotype and parental transmission in myotonic dystrophy. Am J Hum Genet 52: 1164–1174

Harper PS (1989) Myotonic dystrophy, 2d ed. WB Saunders, London and Philadelphia

Harris S, Moncrieff C, Johnson K (1996) Myotonic dystrophy: will the real gene please step forward! Hum Mol Genet 5: 1417–1423

Hawley ME, Kidd KK (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86:409–411

Hill A (1994) Late Miocene and early Pliocene hominoids from Africa. In: Corruccini RS, Ciochon RL (eds) Integrative paths to the past: paleoanthropological advances in honor of F. Clarke Howell. Prentice Hall, Englewood Cliffs, NJ, pp 123–145

Hunter A, Tsilfidis C, Mettler G, Jacob P, Mahadevan M, Surh L, Korneluk R, et al (1992) The correlation of age of onset with CTG trinucleotide repeat amplification in myotonic dystrophy. J Med Genet 29:774–779

Imbert G, Kretz C, Johnson K, Mandel J-L (1993) Origin of the expansion mutation in myotonic dystrophy. Nat Genet 4:72–76

Jansen G, Bachner D, Coerwinkel M, Wormskamp N, Hameister H, Wieringa B (1995) Structural organization and developmental expression pattern of the mouse WD-repeat gene DMR-N9 immediately upstream of the myotonic dystrophy locus. Hum Mol Genet 4:843–852

Jansen G, Groenen PJ, Bachner D, Jap PH, Coerwinkel M, Oerlemans F, van den Broek W, et al (1996) Abnormal myotonic dystrophy protein kinase levels produce only mild myopathy in mice. Nat Genet 13:316–324

Kidd KK, Morar B, Castiglione CM, Zhao H, Pakstis AJ, Speed WC, Bonne-Tamir B, et al. A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. Hum Genet (in press)

Kidd KK, Pakstis AJ, Castiglione CM, Kidd JR, Speed WC, Goldman D, Knowler WC, et al (1996) DRD2 haplotypes containing the TaqI A1 allele: implications for alcoholism research. Alcohol Clin Exp Res 20:697–705

Korczyn AD (1994) Neurologic genetic diseases of Jewish people. Biomed Pharmacother 48:391–397

Krahe R, Ashizawa T, Abbruzzese C, Roeder E, Carango P, Giacanelli M, Funanage VL, et al (1995*a*) Effect of myotonic dystrophy trinucleotide repeat expansion on DMPK transcription and processing. Genomics 28:1–14

Krahe R, Eckhart M, Ogunniyi AO, Osuntokun BO, Siciliano MJ, Ashizawa T (1995*b*) De novo myotonic dystrophy mutation in a Nigerian kindred. Am J Hum Genet 56: 1067–1074

Lewontin RC (1964) The interaction of selection and linkage. I. General considerations: heterotic models. Genetics 49: 49–67

Lichter JB, Barr CL, Kennedy JL, van Tol HHM, Kidd KK, Livak KJ (1993) A hypervariable segment in the human dopamine receptor D4 (DRD4) gene. Hum Mol Genet 2: 767–773

Lotz BP, van den Meyden CH (1985) Myotonic dystrophy. I. A genealogical study in the northern Transvaal. S Afr Med J 67:812–814

Mahadevan MS, Foitzik MA, Surh LC, Korneluk RG (1993) Characterization and polymerase chain reaction (PCR) detection of an Alu deletion polymorphism in total linkage disequilibrium with myotonic dystrophy. Genomics 15: 446–448

Mahadevan MS, Tsilfidis C, Sabourin L, Shutler G, Amemiya C, Jansen G, Neville C, et al (1992) Myotonic dystrophy mutation: an unstable CTG repeat in the 3′ untranslated region of the gene. Science 255:1253–1255

Mor-Cohen R, Magal N, Gadoth N, Achiron A, Shohat T, Shohat M (1997*a*) Correlation between the incidence of myotonic dystrophy in different groups in Israel and the number of CTG trinucleotide repeats in the myotonic gene. Am J Med Genet 71:156–159

——— (1997*b*) The lower incidence of myotonic dystrophy in Ashkenazic Jews compared to North African Jews is associated with a significantly lower number of CTG trinucleotide repeats. Isr J Med Sci 33:190–193

Mulley JC, Staples A, Donnelly A, Gedeon AK, Hecht BK, Nicholson GA, Haan EA, et al (1993) Explanation for exclusive maternal origin for congenital form of myotonic dystrophy. Lancet 341:236–237

Neville CE, Mahadevan MS, Barceló JM, Korneluk RG (1994) High resolution genetic analysis suggests one predisposing haplotype for the origin of the myotonic dystrophy mutation. Hum Mol Genet 3:45–51

Primmer CR, Ellegren H, Saino N, Moller AP (1996) Directional evolution in germline microsatellite mutations. Nat Genet 13:391–393

Rubinsztein DC, Leggo J, Amos W, Barton DE, Ferguson-Smith MA (1994) Myotonic dystrophy CTG repeats and the associated insertion/deletion polymorphism in human and primate populations. Hum Mol Genet 3:2031–2035

Shriver MD, Jin L, Chakraborty R, Boerwinkle E (1993) VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics 134:983–993

Stringer CB, Andrews P (1988) Genetic and fossil evidence for the origin of modern humans. Science 239:1263–1268

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K-H, Bonne-Tamir B, et al (1996*a*) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380–1387

Tishkoff SA, Ruano R, Kidd JR, Kidd KK (1996*b*) Distribution and frequency of a polymorphic *Alu* insertion at the plasminogen activator locus in humans. Hum Genet 97:759–764

Tsilfidis C, MacKenzie AE, Mettler G, Barceló J, Korneluk RG (1992) Correlation between CTG trinucleotide repeat length and frequency of severe congenital myotonic dystrophy. Nat Genet 1:192–195

Valdes AM, Slatkin NB, Freimer NB (1993) Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics 133:737–749

Wang YH, Griffith J (1995) Expanded CTG triplet blocks from the myotonic dystrophy gene create the strongest known natural nucleosome positioning elements. Genomics 25: 570–573

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2:1123–1128

Weir BS (1996) Genetic data analysis II. Sinauer, Sunderland, MA

Wilson AC, Sarich VM (1969) A molecular time scale for human evolution. Proc Natl Acad Sci USA 63:1088–1093

Yamagata H, Miki T, Nakagawa M, Johnson K, Deka R, Ogihara T (1996) Association of CTG repeats and the 1-kb Alu insertion/deletion polymorphism at the myotonin protein kinase gene in the Japanese population suggests a common Eurasian origin of the myotonic dystrophy mutation. Hum Genet 97:145–147

Yamagata H, Miki T, Ogihara T, Nakagawa M, Higuchi I, Osame M, Shelbourne P, et al (1992) Expansion of unstable DNA region in Japanese myotonic dystrophy patients. Lancet 339:692

Zerylnick C, Torroni A, Sherman SL, Warren ST (1995) Normal variation at the myotonic dystrophy locus in global human populations. Am J Hum Genet 56:123–130

Zhang L, Leeflang EP, Yu J, Arnheim N (1994) Studying human mutations by sperm typing: instability of CAG trinucleotide repeats in the human androgen receptor gene. Nat Genet 7:531–535